

Characterizing the Topography of Multi-dimensional Energy Landscapes

H. Lydia Deng
Landmark Graphics Corp
Highlands Ranch, Colorado USA

John A. Scales
Department of Physics
Colorado School of Mines
Golden, Colorado 80401, USA

A basic issue in optimization, inverse theory, neural networks, computational chemistry and many other problems is the geometrical characterization of high dimensional functions. In inverse calculations one aims to characterize the set of models that fit the data (among other constraints). If the data misfit function is unimodal then one can find its peak by local optimization methods and characterize its width (related to the range of data-fitting models) by estimating derivatives at this peak. On the other hand, if there are local extrema, then a number of interesting and difficult problems arise. Are the local extrema important compared to the global or can they be eliminated (e.g., by smoothing) without significant loss of information? Is there a sufficiently small number of local extrema that they can be enumerated via local optimization? What are the basins of attraction of these local extrema? Can two extrema be joined by a path that never goes uphill? Can the whole problem be reduced to one of enumerating the local extrema and their basins of attraction? For locally ill-conditioned functions, premature convergence of local optimization can be confused with the presense of local extrema. Addressing any of these issues requires topographic information about the functions under study. But in many applications these functions may have hundreds or thousands of variables and can only be evaluated pointwise (by some numerical method for instance). In this paper we describe systematic (but generic) methods of analysing the topography of high dimensional functions using local optimization methods applied to randomly chosen starting models. We provide a number of quantitative measures of function topography that have proven to be useful in practical problems along with error estimates.

PACS numbers: 91.30, 02.70, 02.50, 02.50.N

WHAT MAKES AN OPTIMIZATION PROBLEM HARD?

We consider the problem of optimizing a function F (the *objective* or *cost* function) mapping $\mathcal{M} \subset \mathbf{R}^N$ into $\mathcal{Y} \subset \mathbf{R}$. We refer to \mathcal{M} as the *model space*, and each point in the model space, \mathbf{m} , is a *model*. Depending on the application, the goal may be to find the global extremum of F , a single local extremum, or a collection of local extrema. In this paper we will assume that optimization refers to minimization, whether local or global.

There is no generally agreed upon characterization of what makes an optimization problem hard. Hardness has to do partly with our goals — do we need a global extremum or will a good local extremum do; partly with the structure of the function — does it have lots of local extrema, how broad are the basins associated with these extrema; and partly with the dimensionality of the problem — exhaustive search will be infeasible except for low-dimensional problems.

In many applications, however, the function F cannot be expressed in closed form in terms of elementary functions, but can only be evaluated point-wise by computer programs. Such problems arise in many fields. Some of the most widely studied include the *spin-glass* problem, the *traveling-salesman* problem (TSP), and the *residual statics* problem of exploration seismology.

Global Search Strategies

If the structure of function F is unknown, optimization is fundamentally a matter of search in the model space. In order to be able to treat such a broad variety of situations, we begin with an abstract statement of a search algorithm. Here, we use the notation $\tilde{\mathbf{m}}^t$ to represent a population of candidate models at the time step t .

Algorithm 1 General Search (GS) $\tilde{\mathbf{m}} = GS(F, P, \mathbf{T}, S)$

Let $F : \mathcal{M} \subset \mathbf{R}^N \rightarrow \mathcal{Y} \subset \mathbf{R}$, $P \equiv \tilde{\mathbf{m}}^0 = \{\mathbf{m}_k^0\}_{k=1, \dots, K}$ be an initial population of models, where $\mathbf{m}_k^0 \in \mathcal{M}$ and $K \geq 1$, \mathbf{T} a transition operator, and S a stopping criterion.

1. Iteratively apply the transition operator to generate a new population of models at each iteration, so that $\tilde{\mathbf{m}}^t = \mathbf{T} \tilde{\mathbf{m}}^{t-1}$;
2. Repeat (1) until S is satisfied. The final set of models $\tilde{\mathbf{m}}$ are the output of the search.

Any searching process can be considered as an evolution of a population of models (possibly a single model) in the N -dimensional model space. The transition operator \mathbf{T} is the rule that determines to which models the population evolves from the previous population. Here, we assume that the transition operator \mathbf{T} is independent of the time step t , which is the case in most algorithms. Different optimization algorithms differ by the strategies in choosing the initial population P and the rules of transition from one population of models to another, \mathbf{T} .

Among the searching methods defined via Algorithm 1, there are two extreme strategies, *hill-climbing* (HC) and *uniform Monte Carlo* (UMC). HC search is a local descent search applied to a single model (population size $K = 1$). An initial model $P = \mathbf{m}^0$ is selected (possibly at random) and the transition operators $\mathbf{T} = \mathbf{T}_{local}$ are deterministic operators, such as conjugate gradient, quasi-Newton, or downhill simplex, which follow a path downhill as far as possible. For objective functions containing more than one local extremum (*multi-modal*), the result of HC strongly depends on the choice of the initial model \mathbf{m}^0 . UMC, on the other hand, selects points with uniform probability in the model space. The transition operation \mathbf{T} is simply the selection of new points at random and therefore makes no use of information from previous generations. Thus, if there are N parameters and each of them can take m possible values, the probability of finding a particular model is proportional to m^{-N} for each function evaluation using UMC.

Search strategies have been developed that yield a compromise between these two extremes; almost all of these incorporate stochastic elements, especially in the construction of transition operators. It is important for the success of global searches that the transition operators make the best use of information provided by the current samples while avoiding being trapped in local extrema. Among all these strategies, the most widely used are *Simulated Annealing* (SA) [18], *Genetic Algorithms* (GA) [15] and random hill-climbing (RHC), to be defined shortly.

SA and GA searching strategies use stochastic transition operators \mathbf{T} that are biased towards good samples from the previous generations. Many variations of SA and GA can be found in the literature [1, 13, 27]. Although the asymptotic convergence results are known for both SA [14] and GA [6] these results are hardly useful in practice.

RHC searches, on the other hand, apply deterministic transition operations \mathbf{T} to a randomly chosen population P . Hence RHC explores locally in multiple areas of objective functions, and the resulting samples are a set of local/global extrema. This search algorithm can be described as

Algorithm 2 Random Hill Climbing $\tilde{\mathbf{m}} = RHC(F, K, \epsilon, cmax)$

Let the randomly chosen initial population size be K . Let the stopping criterion S be that either gradients of all samples are reduced to the tolerance ϵ or the number of iterations reaches a maximum $cmax$. Let \mathbf{T}_{local} be a local descent search operator.

1. Choose initial models $P = \{\mathbf{m}_k^0\}_{k=1, \dots, K} \in \mathcal{M}$ uniformly at random, where $K \gg 1$;
2. Apply Algorithm 1, $\tilde{\mathbf{m}} = GS(F, P, \mathbf{T}_{local}, S)$.

The final population contains n distinct models, $\tilde{\mathbf{m}} = \{\mathbf{m}_k\}_{k=1, \dots, n}$.

By *uniformly random* we mean that each components of the initial models are chosen randomly with uniformly probability between the maximum and minimum possible values. In this paper, all RHC numerical results use non-linear Conjugate Gradient as transition operators [8].

Landscape of Objective Functions

Chavent [5] developed sufficient conditions for an objective function to be locally convex. These conditions are based on the distance \times curvature induced by the objective function on trajectories. In principle, this local convexity criterion could be generalized to global samples of an objective function, to provide a global measure of complexity.

On the other hand, imagine the surface of an objective function being a high-dimensional landscape with hills and basins of different depths and widths scattered on the surface. Performance of searching algorithms depends to a large extent on topographical features on this landscape.

For SA and GAs, this situation is summarized heuristically by Kaufmann [17]:

“Annealing works well only in landscapes in which deep energy wells also drain wide basins. It does not work well on either a random landscape or a “golf course” potential, which is flat everywhere except for a unique “hole”. In the latter case, the landscape offers no clue to guide search.

Recombination (in GAs) is useless on uncorrelated landscapes but useful under two conditions (1) when the high peaks are near one another and hence carry mutual information about their joint locations in genotype space and (2) when parts of the evolving system are quasi-independent of one another and hence can be interchanged with modest chances that the recombined system had the advantage of both parents.”

In addition, since RHC uses local-descent transition operators, its performance will also be strongly influenced by topography.

It has been proposed that functions can be characterized by their spatial correlation properties [23, 30]. Several typical combinatorial optimization problems have been investigated by studying the correlation in landscapes: the TSP [25], graph-bipartitioning problem [24], and the NK model problems, a spin-glass like problem in biology [16]. Using correlation features of the objective function’s landscape as a criterion, these authors study the effectiveness of particular global algorithms for certain types of landscapes.

In addition, analyzing the topography of high-dimensional energy functions is important in physics and chemistry. Berry and Breitengraser-Kunz [3] studied topography and dynamics of multidimensional inter-atomic potential surfaces by analyzing a population of local minima, each of which has two saddle points connected to it. By connecting these samples in a certain order, the high-dimensional function surface is represented by a series of one-dimensional lines. By looking at these one-dimensional plots, the topography information is represented by the width and depth of the primary, secondary or tertiary basins of attractions [3].

The structure of high-dimensional Hamiltonians has also been studied by means of entropy [10]. For an N -dimensional Hamiltonian, a collection of local extrema are first found by some means. Contributions of these local-minima are represented by a probability distribution $\{p_i\}_{i=1, \dots, n}$ where

$$p_i \propto \Delta^{(i)}(N) = \prod_{k=1}^N \delta_k^i. \quad (1)$$

Here, δ_k^i is the estimated width of the i th basin of attraction along the k th coordinate. The N -dimensional surface is then characterized by the following entropy,

$$S(N) = - \sum_i p_i \ln p_i = \left\langle \ln \left(\frac{1}{\Delta(N)} \right) \right\rangle. \quad (2)$$

In this paper, we use a similar measure. However, we estimate p_i by random hill-climbing rather than equation 1. Further, we base our measure not on p_i itself, but rather on a related probability that takes into account of the values of the local minima. We also perform a confidence interval analysis. Finally, as a concrete application, we show that this measure can be used to compute the optimal simplification of a multi-resolution analysis (MRA) of highly non-convex seismic optimization problem.

MEASURES OF TOPOGRAPHY

Definitions

The surface topography of functions is largely associated with the number of local minima, widths of the basin of attractions associated with these minima, and relative depths of these basins. The *basin of attraction* associated with the i th local minimum may be loosely defined as *the maximum volume A_i in the N -dimensional model space within which all models can converge to the i th local minima after infinite number of iterations by a local descent search algorithm*. Suppose the volume of the entire model space is represented as M , then the ratio $p_i = \frac{A_i}{M}$ is the probability of converging to the i th local minimum for a uniformly random model. The following definition serves to introduce three quantitative measures of topography: a probability associated with the relative volumes of the basins of attraction (p), a version of p scaled by the estimated depths of the basins of attractions (q) and the entropy of q .

Definition 1 Entropy-Based Topography

Let $F : \mathcal{M} \subset \mathbf{R}^N \rightarrow \mathcal{Y} \subset \mathbf{R}$ be bounded and have n isolated local minima, where n is finite. Let $\{\mathbf{m}_i\}_{i=1,\dots,n}$ be these distinct local minima, and $\{y_i = F(\mathbf{m}_i)\}_{i=1,\dots,n}$ be their corresponding function values. Let p_i be the probability that a model chosen with uniform probability in \mathcal{M} will converge to the i th local minimum under the action of an exact local optimization algorithm.

Define $\{q_i\}_{i=1,\dots,n}$ to be a probability distribution where

$$q_i \propto \begin{cases} p_i, & \text{if } \sigma = 0; \\ p_i e^{-\frac{|y_i - y_m|}{\sigma}}, & \text{otherwise,} \end{cases} \quad (3)$$

where $i \in [1, n]$, y_m is the value at the global minimum, $\sigma = \frac{1}{n} \sum_{j=1}^n |y_j - y_m|$, and $\sum_{i=1}^n q_i = 1$. The entropy is defined to be,

$$\begin{aligned} C_e &= - \sum_{i=1}^n q_i \ln(q_i) \\ &= \left\langle \ln \left(\frac{1}{q_i} \right) \right\rangle, \end{aligned} \quad (4)$$

where the angle brackets denote the expected value with respect to the probability distribution $\{q_i\}_{i=1,\dots,n}$.

For the entropy in Definition 1, it is always the case that $C_e \geq 0$. If the function is unimodal (only one extremum), then $C_e = 0$. On the other hand, if the function has n isolated and equally valued local extrema, then C_e is $\ln n$. Therefore, the entropy increases with the number of local extrema n .

As a simple example, Figure 1 shows two one-dimensional functions with the same number of local minima and widths of basins of attractions ($A_i \propto p_i$). However, the difficulty of minimizing these functions is different: the left function has identical basins of attractions, while the one on the right has a dominant global minimum at $x = 0$ and decreasingly important local minima away from the center. The entropy of Definition 1 gives a higher C_e value to that of the function on the left ($C_e = 2.2$) than that on the right ($C_e = 1.5$).

Since the functions we are interested in can usually be evaluated only point-wise, the number of local minima n and $\{p_i\}_{i=1,\dots,n}$ are not known. Some degree of global sampling is essential in order to achieve the characterization we seek. As shown in Algorithm 2, RHC explores various regions of the model space and takes initial samples down-hill to the bottom of the basins on the surface of functions. Therefore, a statistical analysis of results of systematic RHC searches can be used to estimate the topographic quantities.

Suppose the local-descent search is ideal, i.e. all initial models converge to exact local minima, the number of models converging to each local minima from K randomly chosen initial models has a multinomial probability distribution. If the initial models are randomly chosen under a uniform probability distribution, the probability of converging to the i th local minimum is proportional to the width of the i th basin of attraction, p_i . Let K_1, \dots, K_n be the random

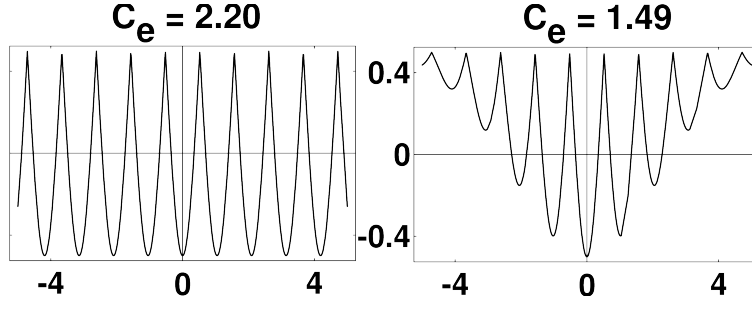


FIG. 1. Two one-dimensional functions with the same number and widths of basins of attractions. The function on the left have entropy $C_e = 2.2$, while the function on the right has $C_e = 1.5$.

variables representing the frequency of models converging to each of the local minima. For the population of K , the joint probability density of these random variables is

$$f(k_1, \dots, k_n) = \frac{K!}{k_1! \dots k_n!} p_1^{k_1} \dots p_n^{k_n}, \quad (5)$$

where $\sum_{i=1}^n k_i = K$. For each $i \in [1, n]$, the mean value of the random variable k_i is $E[k_i] = K p_i$. Therefore, the convergence frequency of a RHC of large population can be used to estimate the number of local minima as well as the widths of basin of attractions. Hence, the estimation of the entropy measure in Definition 1 can be defined as follows.

Definition 2 Entropy-Based Estimates

Let $F : \mathcal{M} \subset \mathbf{R}^N \rightarrow \mathcal{Y} \subset \mathbf{R}$ be bounded and have finite number of isolated local minima. Let $\{\mathbf{m}_i\}_{i=1, \dots, \hat{n}} = \mathbf{RHC}(F, K, \epsilon, cmax)$ be the distinct converged models of RHC searches. Let $\{k_i\}_{i=1, \dots, \hat{n}}$ be the frequency distribution of the final population, and $\{y_i = F(\mathbf{m}_i)\}_{i=1, \dots, \hat{n}}$ be their corresponding function values.

Define the estimated entropy \hat{C}_e as

$$\begin{aligned} \hat{C}_e &= - \sum_{i=1}^{\hat{n}} \hat{q}_i \ln(\hat{q}_i) \\ &= \left\langle \ln \left(\frac{1}{\hat{q}_i} \right) \right\rangle, \end{aligned} \quad (6)$$

where \hat{q}_i is normalized to a probability distribution $\sum_{i=1}^{\hat{n}} \hat{q}_i = 1$, and $\hat{q}_i \propto x_i v_i$, in which

$$x_i \equiv \frac{k_i}{K}, \quad (7)$$

and

$$v_i \equiv \begin{cases} 1, & \text{if } \sigma = 0; \\ \exp(-\frac{|y_i - y_m|}{\sigma}), & \text{otherwise,} \end{cases} \quad (8)$$

where $y_m = \min \{y_i\}$ and $\sigma = \frac{1}{\hat{n}} \sum_{j=1}^{\hat{n}} |y_j - y_m|$.

Definition 2 is a statistical estimation of the entropy in Definition 1. The exact entropy C_e characterizes topographical features of objective function, and hence independent of numerical computation and any searching technique. The estimation \hat{C}_e , however, would be influenced by numerical issues. If, for examples, the curvature of the function is nearly zero, which is equivalent to an ill-conditioned Hessian matrix, gradient-based local descent searches may not converge to the exact local minima. The estimated value of \hat{C}_e in such a situation may be higher than the true complexity C_e . In practice, however, it is often difficult to distinguish the results of such ill-conditioning from those of multi-modality. Therefore, taking such numerical issues into account can represent an important aspect in the difficulty of optimization.

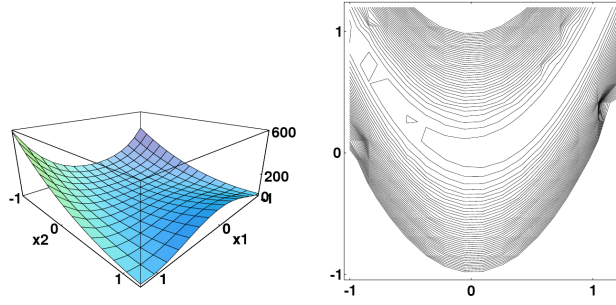


FIG. 2. Two-dimensional Rosenbrock function. The figure on the left is a 3-D plot of the function surface, while the one on the right shows the contour plot of the same function.

Numerical Examples

In this section, we use the entropy \hat{C}_e to study two commonly used test functions in optimization, the Rosenbrock and Griewank functions.

N-dimensional Rosenbrock function

An N -dimensional Rosenbrock function can be written as

$$R(\mathbf{x}) = \sum_{i=1}^{N-1} [100(x_i - x_{i-1}^2)^2 + (1 - x_{i-1})^2], \quad (9)$$

where $\mathbf{x} = (x_0, \dots, x_N)$. Although unimodal, the long and narrow basin is a challenge for searching algorithms. Figure 2 shows the function surface and its contour when $N = 2$. When $N \geq 2$, the function is still unimodal, but it is not easy to see how the increase of dimensionality alters the difficulty of optimization.

One way of studying the spatial curvature of functions is by looking at the ratio of largest and smallest eigenvalues (*condition number*) of the Hessian at a point. The Hessian for equation (9) is a tri-diagonal matrix,

$$\begin{pmatrix} a_0 & c_0 & 0 & \cdots & 0 \\ b_1 & a_1 & c_1 & 0 & \cdots \\ 0 & b_2 & a_2 & c_2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & b_{N-1} & a_{N-1} \end{pmatrix} \quad (10)$$

where

$$a_i = \begin{cases} 2 + 1200x_0^2 - 400x_1, & \text{if } i = 0; \\ 202 + 1200x_{i-1}^2 - 400x_i, & \text{if } 0 < i < N-1; \\ 200, & \text{if } i = N-1, \end{cases}$$

$$\begin{aligned} b_i &= -400x_{i-1}, & 0 < i \leq N-1, \\ c_i &= -400x_{i+1}, & 0 \leq i < N-1. \end{aligned}$$

At the global minimum $(1, 1, \dots, 1)$, the tri-diagonal matrix equation (10) becomes Toeplitz except for a_0 and a_{N-1} . The condition number of the Hessian at the global minimum reaches an asymptote with increasing dimension, as shown in Figure 3. Figure 4 shows \hat{C}_e as a function of the number of dimensions; it shows the same asymptotic trend as does the condition number. Thus the increasing complexity for low dimensions is the result of increasing ill-conditioning of the Hessian and has nothing to do with local minima.

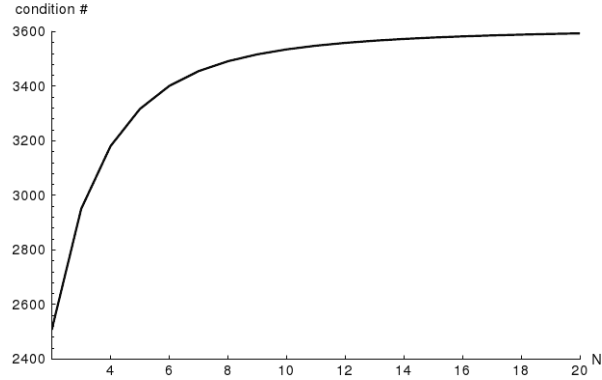


FIG. 3. Condition number of the Hessian matrix for the N-dimensional Rosenbrock function at the global minimum.

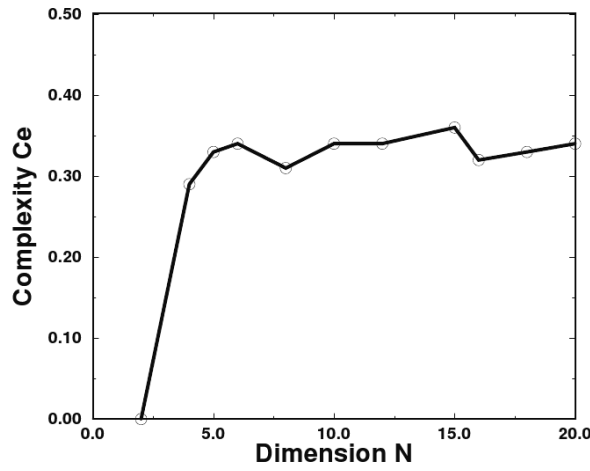


FIG. 4. \hat{C}_e for N-dimensional Rosenbrock functions as a function of N .

High-dimensional Griewank functions

The Griewank function is also used to test optimization algorithms [26, 31]:

$$g(\mathbf{x}) = 1 + \sum_{i=1}^N \frac{x_i^2}{4000} - \prod_{i=1}^N \cos\left(\frac{x_i}{\sqrt{i}}\right) \quad (11)$$

The cosine term makes equation (11) multi-modal. Figure 5 shows a one-dimensional slice of the Griewank function along the diagonal of the hypercube for dimensions 1, 3, 5, 9. Whitley et al. [31] observed such slices and concluded that “as the dimensionality increases the local optima induced by the cosine decrease in number and complexity”.

However, such pictures can be misleading since they tell us only about low-dimensional projections of the function. Figure 6 shows slices of the same functions when all but one variables are fixed to be 0. The increasing dimensionality does not change the oscillation around the global minimum at the origin. Therefore, studying the overall performance of high dimensional functions could be tricky. We compute \hat{C}_e for the Griewank function with a population 500 and 1000 models in the hyper-cube of $-10 \leq x_i \leq 10$, $i = 0, \dots, N-1$. Figure 7 shows the resulting \hat{C}_e for dimensions up to 50 for initial populations of both 500 and 1000. Both curves in Figure 7 give us consistent results that the complexity of Griewank function in this range increases till dimension around 9, then decreases when number of dimension continuous to increase. This result can be verified by the analysis of Griewank function. Therefore, using the entropy we can understand more comprehensively the dimensional-dependence of complexity of certain functions than by simply looking at hyper-planes.

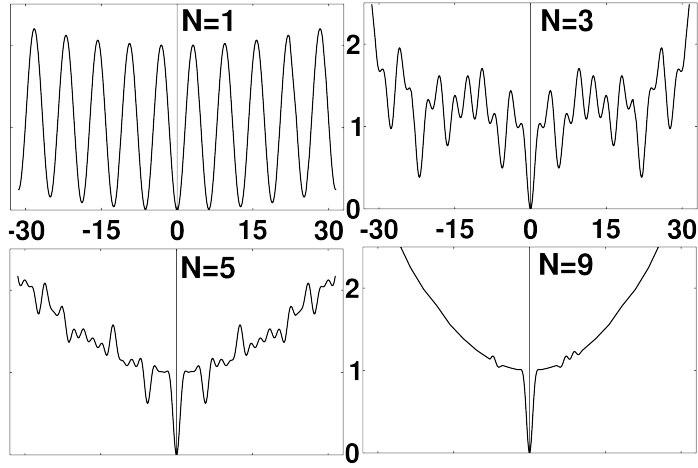


FIG. 5. Diagonal slices of N -dimensional Griewank functions.

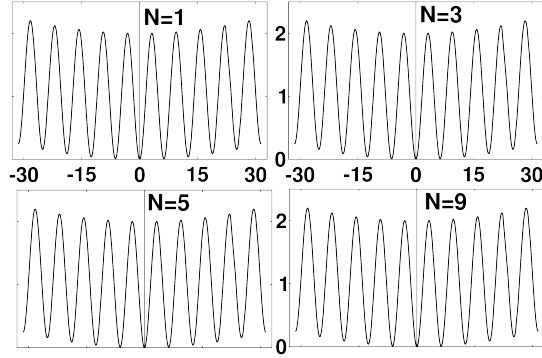


FIG. 6. Slices of N -dimensional Griewank functions. All variables but one are fixed at 0.

CONFIDENCE INTERVALS ANALYSIS

Next we derive confidence intervals on the entropy in Definition 2. The following analysis is based on the assumption of *ideal RHC*, which is a special case of Algorithm 2 where an infinitely large $cmax$ is allowed and ϵ is infinitely small. First, it is easy to prove that as long as the population size K is large enough, x_i defined in equation (7) would be good approximation to p_i for $\forall i \in [1, n]$. We have the following theorem, the proof of which is given in the appendix.

Theorem 1 *Let $F : \mathcal{M} \subset \mathbf{R}^N \rightarrow \mathcal{Y} \subset \mathbf{R}$ be bounded and have finite number of isolated local minima. Let p_i be the probability of converging to the i th local minimum for a starting model chosen with uniform probability on \mathcal{M} . Perform an ideal RHC as defined in Algorithm 2 with an initial population of K . Let α and β be related by the following equation,*

$$Pr(|z| \leq \beta) = 1 - \alpha, \quad (12)$$

where z has a standard-normal distribution $N(0,1)$.

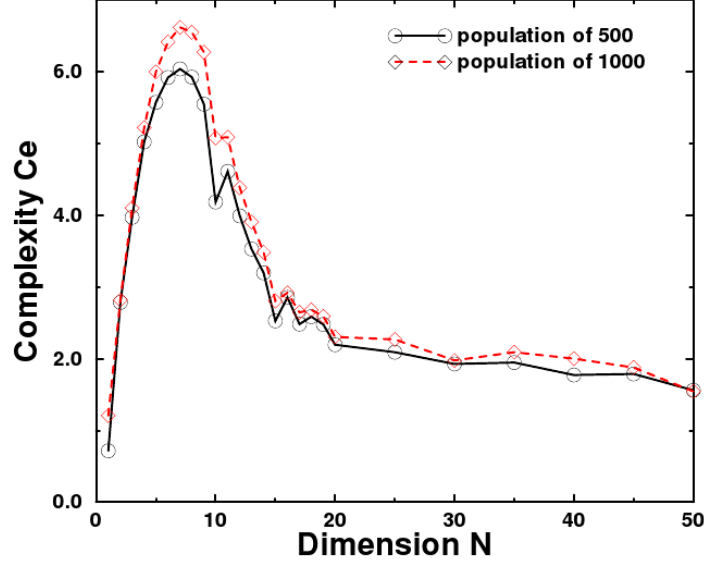


FIG. 7. \hat{C}_e as a function of dimension N for the Griewank function with populations of 500 and 1000.

Let x_i be as defined in equation (7). If the population K is such that $K p_i \geq 5$ for any $i \in [1, n]$, we have the following,

1. x_i has an approximate normal distribution with

$$E[x_i] = p_i, \quad (13)$$

and

$$\text{var}(x_i) = \frac{p_i(1-p_i)}{K}. \quad (14)$$

2. x_i is an unbiased, consistent estimator of p_i .
3. With confidence of $(1-\alpha)\%$, the error associated with estimating p_i from x_i is bounded by

$$\beta \sqrt{\frac{x_i(1-x_i)}{K}}. \quad (15)$$

To get some ideas of the magnitudes of the population size and the confidence interval, here is a simple example.

Example 1 If for a problem as described in Definition 2, we have $p_i = 0.01$. Then for approximating the binomial distribution with a normal distribution, we need at least $K > 500$.

Example 2 For the same problem as stated in Example 1, suppose a population size of 520 was used in an ideal RHC, and 100 of the models converged to the i th local minimum. Then, $x_i \approx 0.192$. If we want to have 90% confidence, then $\beta = 1.65$. The error bound for the estimation of p_i with x_i would be 0.0285. That is, with 90% confidence, we can say that $0.175 \leq p_i \leq 0.220$. If, on the other hand, we want to have 95% confidence for this estimation when $\beta = 1.96$, then $0.158 \leq p_i \leq 0.226$.

In the following theorem, we estimate the distribution and error bound of the estimation for the entropy Definition 2. The proof of the following theorem is also given in the appendix.

Theorem 2 Let $F : \mathcal{M} \subset \mathbf{R}^N \rightarrow \mathcal{Y} \subset \mathbf{R}$ be bounded and have finite number of isolated local minima. Let $p = \{p_i\}_{i=1,\dots,n}$ be the probability distribution in Definition 1, and $p_m = \min\{p_i\}$. Let C_e be the entropy of F (as in Definition 1) and suppose the RHC of population K is ideal. As a result, the initial population of models converge to different local extrema with a frequency distribution of $\{k_i\}_{i=1,\dots,\hat{n}}$. Finally, let x_i be defined as in equation 7, and $\hat{C}(x_1, x_2, \dots, x_{\hat{n}})$ the estimated entropy (as in Definition 2). If α and β are defined as in equation (12), then we can prove the following statements:

1. \hat{C}_e has an approximate normal distribution with

$$E[\hat{C}_e] = C_e, \quad (16)$$

and

$$\text{var}(\hat{C}_e) = \sum_{i=1}^n \frac{c v_i (1 + \ln q_i)^2}{K} q_i - \frac{(1 - C_e)^2}{K}, \quad (17)$$

in which

$$q_i = c v_i p_i, \quad (18)$$

for $\forall i \in [1, n]$, c is a scale factor so that $\sum_{i=1}^n q_i = 1$, and v_i is as defined in equation (8).

2. \hat{C}_e is an unbiased, consistent estimator of C_e .
3. Let $\hat{q}_{min} = \min_{i \in [1, \hat{n}]} \{\hat{q}_i\}$. If the population size K is such that $K p_m \geq 5$, then with confidence of $(1 - \alpha) 100\%$, the estimation error of the complexity is at most $\beta \delta$ where $\delta \geq 0$. That is,

$$|C_e - \hat{C}_e(x_1, \dots, x_{\hat{n}})| \leq \beta \delta,$$

where

$$\delta^2 = \frac{c}{K} (1 - (2 + \ln \hat{q}_{min}) \hat{C}_e) - \frac{(1 - \hat{C}_e)^2}{K} > 0. \quad (19)$$

Remark 1 When the number of local minima, n , is large, the real p_m may very small. Then, an unrealistically large initial population size K may be required to satisfy $K p_m \geq 5$. Realistically, we have to content ourselves with not being able to find all local minima in such difficult situations. If the smallest basin we found with a K -population RHC is \hat{p}_m and $\hat{p}_m > p_m$, there are some $p_i < \{\hat{p}\}_m$ which are not found by the RHC. Their corresponding q_i would not be accounted for in the complexity estimation. However, the contributions of these narrow basins to the complexity are proportional to $q_i \log q_i$. Since, $\lim_{q \rightarrow 0} q_i \log q_i = 0$, the error caused by these narrow basins will be small as long as $q_i = c v_i p_i$ is small. Since $0 < v_i \leq 1$ for $\forall i \in [0, n]$ by definition, these conditions can be easily satisfied as long as these narrow basins are not global minima.

We conclude this section by showing an example of the evaluation of the confidence interval for the Griewank function. It is important to note that in such an analysis, it is assumed that the RHC algorithms are exact. That is, numerical effects are ignored.

Example 3 We want to evaluate the confidence interval for the complexity calculation of the 9-dimensional Griewank function shown in equation (11). In Figure 7, we show that the complexity estimation for the population of $K = 1000$ is 6.27. Using the calculated data, $\ln q_{min} = -13.62$, $c = 1.86$, and $\hat{C}_e = 6.37$, we can estimate that $\delta \approx 0.33$. So, with 90% confidence, the error bound would be ± 0.65 . Therefore, we can say that the true complexity value C_e is between 6.9 and 5.6.

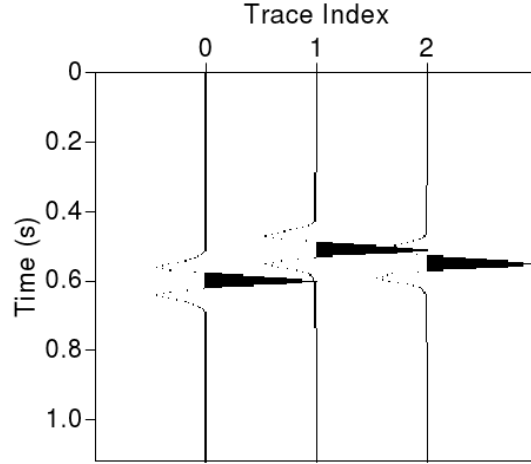


FIG. 8. Three synthetic seismic traces that are shifted by random statics.

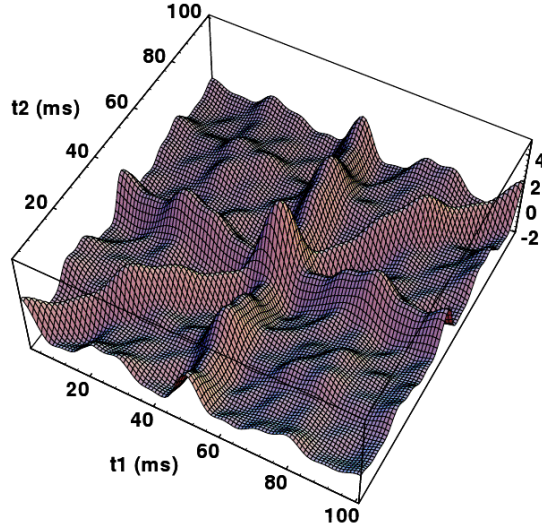


FIG. 9. Landscape of a 2-D residual statics objective function.

A GEOPHYSICAL APPLICATION

Estimating Near-Surface Heterogeneities

In exploration seismology, “statics” are the time shifts in seismic reflection data caused by heterogeneous material properties in the near surface. This causes jitter in the data and degrades processing procedures designed to enhance signal-to-noise, such as averaging. It is possible to formulate an optimization procedure for these static time shifts (the objective function being the power of the averaged data as a function of time shifts), but the resulting optimization problem is highly non-convex [19]. This is illustrated in Figure 9 with a toy example.

Consider an example where we need to align three otherwise identical traces. Fixing the first trace, we look for time-shifts for the second and third traces, t_1, t_2 , so that the sum of squares of the stacked traces (stacking-power) is maximized. Figure 9 shows an example of such a two-dimensional objective function, which has hills and basins of attractions scattered on the landscape. In practice, however, the stacking-power objective function is high-dimensional and highly multi-modal. Monte Carlo global optimization have become an important tool for solving large-scale statics problems [20, 21].

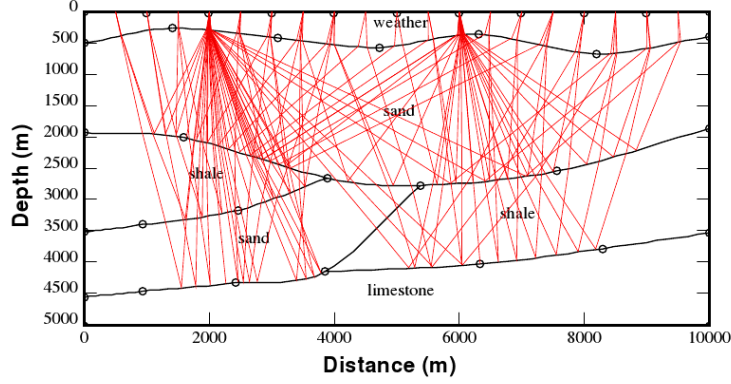


FIG. 10. A hypothetical model of the Earth's upper crust showing rays associated with seismic waves propagating down from sources on the surface, reflecting off geologic boundaries and traveling upwards to receivers on the surface. Static-shifts of the seismic traces are caused by the combined time-distortions of near-source and near-receiver heterogeneities in the weathering layer.

Figure 9 shows a statics objective function with two unknowns. In practice, however, the time-shifts of the traces are not independent. The statics of each trace are caused by the combined time distortion of near-source and near-receiver heterogeneities (*source-statics* and *receiver-statics*). Figure 10 illustrates the similarity of travel paths near each source and each receiver.

The recorded reflection seismic signals are usually sorted into *midpoints* y (of the source and receiver locations) and *offsets* h (half distance between the source and receivers). Letting \mathbf{s} and \mathbf{r} be unknown vectors of source- and receiver-statics, this optimization problem can be formulated as

$$\max_{\mathbf{s}, \mathbf{r}} F(\mathbf{s}, \mathbf{r}) = \sum_{\mathbf{y}} \sum_{\mathbf{h}_1 \neq \mathbf{h}_2} \Phi_{\mathbf{h}_1, \mathbf{h}_2}^{\mathbf{y}}(\tau(\mathbf{s}, \mathbf{r})), \quad (20)$$

where $\Phi_{h_1, h_2}^y(\tau)$ is the cross-correlation between traces (after a correction for propagation effects known as “normal move-out” has been applied) of offsets h_1 and h_2 at midpoint y evaluated at

$$\tau = s_{i(y, h_1)} + r_{j(y, h_1)} - s_{i(y, h_2)} - r_{j(y, h_2)},$$

and $i(y, h)$ and $j(y, h)$ are the source and receiver indices for midpoint y and offset h , respectively. The function $F(\mathbf{s}, \mathbf{r})$ in equation (20) is called the *stacking-power function*.

Figure 11 shows the recording geometry of one example synthetic data set. This data set has 20 sources, 35 distinct receivers and 320 traces. All traces are identical except for random source and receiver statics. These are generated by repeatedly shifting a single trace of field data. Thus, the objective function of equation (20) has 55 unknowns. When there are no statics in the data, the global maximum of the function is at the origin ($\mathbf{s}_i = \mathbf{0}$, $\mathbf{r}_j = \mathbf{0}$). Figure 12 shows an arbitrary 2-D hyper-planes of the stacking-power function along the 10th source and 20 receiver statics.

We have analyzed a realistic synthetic statics problem involving some 320 seismic traces and 55 unknown static time shifts. A hyperplane through the objective function for this problem is shown in figure 12. In addition to simply computing the entropy of this function we will show how the entropy might be used to quantitatively address issues related to the topography of functions.

Behavior of the Multi-Resolution Analysis

Rather than using Monte Carlo global optimization methods to solve the statics problem as in [20, 21], Deng [7] has proposed, without proof, simplifying the optimization via a multi-resolution analysis (MRA) of the seismic traces. The idea is to use a wavelet decomposition to generate successively simpler representations of the seismic data, thereby eliminating progressively more local extrema from the objective function. To be precise, let us define a Multi-Resolution RHC algorithm:

Algorithm 3 MRHC ($\tilde{\mathbf{m}} = \text{MRHC}(F, L, \epsilon, cmax)$)

Let $\{\mathbf{S}_i\}_{i=L, \dots, 0}$ be a sequence of decreasingly smooth operators to be defined below, with \mathbf{S}_0 an identity operator.

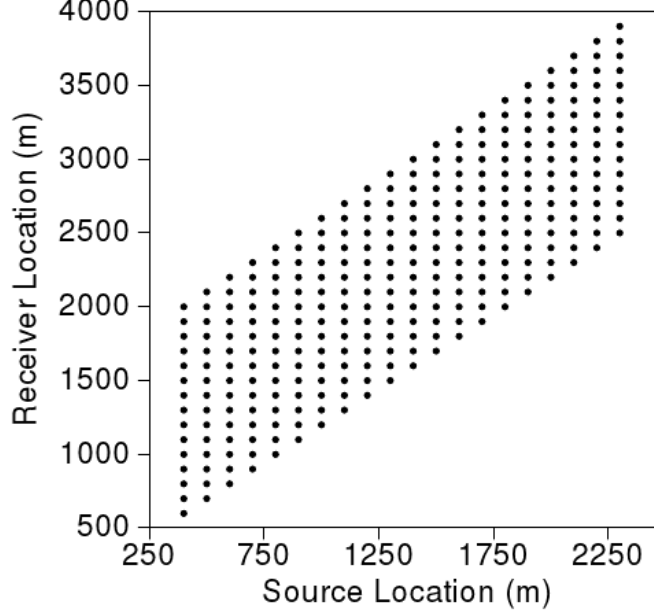


FIG. 11. Recording geometry of a synthetic data set. The horizontal axis is the source position and vertical axis the receiver position.

1. Let $f_L = \mathbf{S}_L F$; choose an initial population $\{\mathbf{m}_k^0\}_{k=1,\dots,K}$ with size K at random; apply Algorithm 2, so $\{\mathbf{m}_k\}_{k=1,\dots,M_L} = RHC(f_L, K, \epsilon, cmax)$, and $i = L - 1$.
2. Let $f_i = \mathbf{S}_i F$, ($L > i \geq 0$) and $\tilde{\mathbf{m}}^0 = \{\mathbf{m}_k\}_{k=1,\dots,M_{i-1}}$; run Algorithm 2, $\{\mathbf{m}_k\}_{k=1,\dots,M_i} = RHC(f_i, K, \epsilon, cmax)$.
3. Decrease the level index i by 1, repeat 2 until $i = 0$. The final set of models $\tilde{\mathbf{m}}$ is the solution.

The smoothing operators $\{\mathbf{S}_i\}_{i=L,\dots,0}$ could be a sequence of low-pass filters with increasingly wider pass-band [4], or a sequence of increasingly fine wavelet operators [7] for decomposing the input seismic data. The sequence of smoothing operators should be such that the resulting functions, $\{f_i\}_{i=L,\dots,0}$, have the same global feature as does the objective function F for all levels and have decreasing number of local optima when the level increases, and $f_0 = F$. Deng [7] showed that this could be achieved using the shift-invariant wavelet basis of Saito and Beylkin [22].

We now apply the entropy-based estimation of complexity to study the multi-resolution analysis of the 55 parameter statics problem introduced in the previous section. Figure 13 shows \hat{C}_e as a function of the wavelet decomposition level using population $K = 1000$; the mean and one-standard deviation error bars are obtained from 32 independent calculations. Results are shown for 6 levels of decomposition using a wavelet operator $\{\mathbf{S}_i\}_{i=0,\dots,5}$ where $i = 0$ is an identity operator, corresponding to use of the original data. These results indicate that for this particular problem a complexity minimum is achieved for a wavelet decomposition of level 4. Higher levels of decomposition actually increase the complexity; presumably this results from the objective function being too flat for local optimization. Thus, the complexity measure gives us a way of choosing a wavelet decomposition level to achieve optimal simplification of an objective function.

CONCLUSIONS

We have developed a collection of simple tools for analysis of the topographic complexity of functions based on the application of local optimization to randomly chosen starting models. In particular we estimate the number of basins of attractions on the function landscape, the widths and depths of these basins and the entropy of the resulting probabilities.

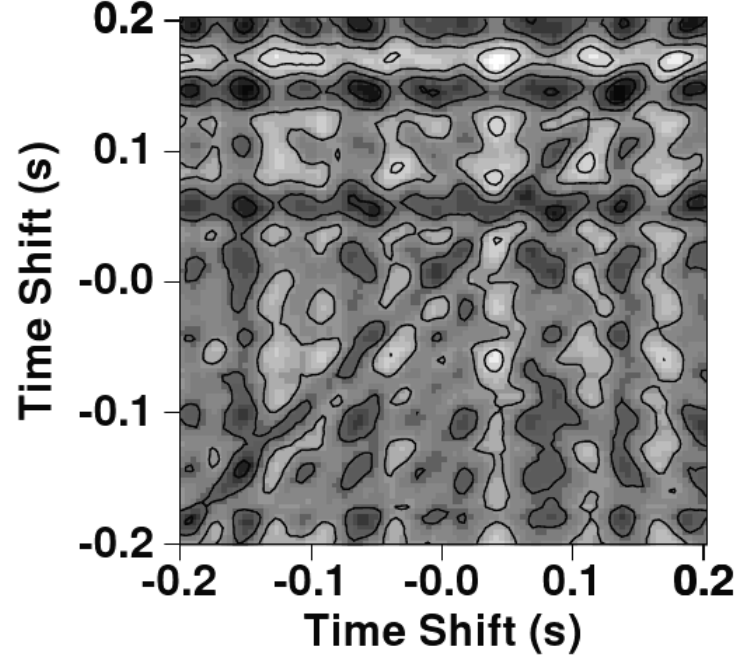


FIG. 12. Hyperplane through the stacking-power function for a problem with 55 unknown static time shifts.

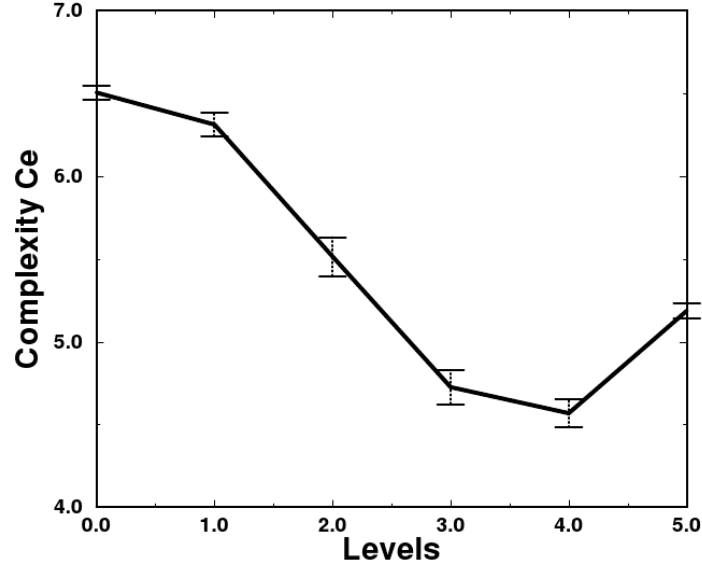


FIG. 13. Complexity \hat{C}_e as a function of the level of wavelet decomposition.

Assuming local descent searches are ideal, we have computed the confidence intervals for the sampling error associated with this complexity measure. There are, on the other hand, several practical issues that we have neglected in this error analysis. Among them, we can mention the convergence error caused by the finite computing time and the finite precision of the local descent algorithms, the criterion for clustering of converged models, and the size of the assumed smallest basin of attraction, p_m . These issues can be investigated by a Monte Carlo analysis as shown in Figure 13.

ACKNOWLEDGMENTS

This work is dedicated to the memory of Albert Tarantola. The authors thank Dr. Bill Navidi for useful discussions and comments on a draft of this work. This work was begun while the authors were at the Center for Wave Phenomena.

-
- [1] E.H.L. Aarts and Jan Korst. *Simulated Annealing and Boltzman Machines*. Wiley, N.Y, 1989.
 - [2] O.M. Becker and M. Karplus. The topography of multidimensional potential energy surfaces: theory and application of peptide structure and kinetics. *Journal of Physical Chemistry*, 106:1495–1517, 1997.
 - [3] R. S. Berry and R. Breitengraser-Kunz. Topography and dynamics of multidimensional interatomic potential surface. *Physical Review Letters*, 74:3951–3954, 1995.
 - [4] C. Bunks, F. M. Saleck, S. Zaleski, and G. Chavent. Multiscale seismic waveform inversion. *Geophysics*, 60(5):1457–1473, September-October 1995.
 - [5] G. Chavent. On the theory and practice of non-linear least-squares. *Adv. Water Resources*, 14:55–63, 1991.
 - [6] T. E. Davis and J. C. Principe. A simulated annealing like convergence theory for the simple genetic algorithm. In R. K. Belew and L.B. Booker, editors, *Proceedings of the fourth international conference on genetic algorithms*. Morgan Kaufmann Publishers, San Mateo, Calif., 1991.
 - [7] H. L. Deng. Using Multi-Resolution Analysis to study the complexity of inverse calculations. *preprint*, 1995.
 - [8] H. L. Deng, W. Gouveia, and J. A. Scales. An object-oriented toolbox for studying optimization problems. In B. H. Jacobsen, K. Moosegard, and P. Sibani, editors, *Inverse Methods, Interdisciplinary Elements of Methodology, Computation, and Applications*, pages 320–330, Berlin, Germany, 1996. Springer-Verlag.
 - [9] J. E. Dennis and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. Prentice-Hall Inc., 1987.
 - [10] M. Falcioni, U. M. B. Marconi, P. M. Ginanneschi, and A. Vulpiani. Complexity of the minimum energy configurations. *Physical Review Letters*, 75:637–640, 1995.
 - [11] R. Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, 1987.
 - [12] J. E. Freund. *Mathematical Statistics*. Prentice Hall, Englewood Cliffs, New Jersey, 5 edition, 1992.
 - [13] D. E. Goldberg and M. P. Samtani. Engineering optimization via genetic algorithms. In *Proceedings of the ninth conference on electronic computation*, pages 471–482. 1986.
 - [14] B. Hajek. Cooling schedules for optimal annealing. *Mathematics of Operations Research*, 13:311–329, 1988.
 - [15] J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Harbor, MI, 1975.
 - [16] S. A. Kauffman and E. D. Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, 141(2):211, 1989.
 - [17] S. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*, chapter 2-3, pages 33–117. Oxford University Press, New York, 1993.
 - [18] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.
 - [19] D. H. Rothman. Nonlinear inversion, statistical mechanics, and residual statics estimation. *Geophysics*, 50:2797–2807, 1985.
 - [20] D. H. Rothman. Nonlinear inversion, statistical mechanics, and residual statics estimation. *Geophysics*, 50:2797–2807, 1985.
 - [21] D. H. Rothman. Automatic estimation of large residual statics corrections. *Geophysics*, 51:332–346, 1986.
 - [22] N. Saito and G. Beylkin. Multiresolution representations using the auto-correlation functions of compactly supported wavelets. *IEEE Transactions on Signal Processing*, 41:3585–3590, 1993.
 - [23] P. F. Stadler. Correlation in landscapes of combinatorial optimization problems. *Europhysics Letters*, 20(6):479–482, Nov 1992.
 - [24] P. F. Stadler and R. Happel. Correlation structure of the landscape of the graph-bipartitioning problem. *Journal of Physics. A*, 25(11):3103–3110, June 1992.
 - [25] P. F. Stadler and W. Schnabl. The landscape of the traveling salesman problem. *Physics Letters A*, 161:337–344, 1992.
 - [26] A. A. Törn and A. Žilinskas. *Global Optimization*. Springer-Verlag, Berlin, Germany, 1989.
 - [27] P.J.M. van Laarhoven and E.H.L. Aarts. *Simulated Annealing: Theory and Practice*. Reidel, Dordrecht, 1987.
 - [28] David J. Wales and Janothan P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *Journal of Physical Chemistry*, pages 5111–5116, 1997.
 - [29] David J. Wales, Mark A. Miller, and Tiffany R. Walsh. Archetypcal energy landscapes. *Nature*, 394:758–760, 1998.
 - [30] E. D. Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biological Cybernetics*, 63:325–336, 1990.
 - [31] D. Whitley, K. Mathias, S. Rana, and J. Dzubera. Building better test functions. San Mateo, Calif., 1995. Morgan Kaufmann Publishers.

Proofs of the Confidence Interval Analysis

Proof of Theorem1

Proof:

Let K_1, \dots, K_n be the random variables that represent the frequency of initial models converging to each local minima. The joint probability density for random variables K_i , $i \in [1, n]$ for population of K is a multinomial distribution. The marginal distribution for each of the random variables is,

$$f(k_i) = \frac{K!}{k_i!(K - k_i)!} p_i^{k_i} (1 - p_i)^{K - k_i}, \quad (21)$$

and the corresponding statistical quantities are,

$$E[k_i] = K p_i, \quad \text{var}(k_i) = K p_i (1 - p_i),$$

for $\forall i \in [1, n]$.

1. For K such that $K p_i \geq 5$, the above binomial distribution can be approximated by a normal distribution. That is, the random variable,

$$\begin{aligned} Z_i &= \frac{k_i - K p_i}{\sqrt{K p_i (1 - p_i)}} \\ &= \frac{x_i - p_i}{\sqrt{p_i (1 - p_i) / K}} \end{aligned}$$

approaches to standard normal distributions $N(0, 1)$ (Theorem 6.8 of [12]), where x_i is defined in equation (7). Therefore, x_i has a normal distribution with the mean and variance as in equations (13) and (14).

2. From equation (13), we see that x_i is an unbiased estimator of p_i . Since $\text{var}(x_i) \propto \frac{1}{K}$, we have

$$\lim_{K \rightarrow \infty} \text{var}(x_i) = 0. \quad (22)$$

Therefore,

$$\lim_{K \rightarrow \infty} x_i = p_i.$$

x_i is also a consistent estimator of p_i for each $i \in [1, n]$.

3. Now with confidence of $(1 - \alpha)100\%$, we have

$$|p_i - x_i| \leq \beta \sqrt{\frac{p_i (1 - p_i)}{K}}, \quad (23)$$

where the value of $\beta = z_{\alpha/2}$ can be looked up from a standard normal distribution table.

However, we do not know p_i in advance. Approximating p_i by x_i when K is large, we have the confidence interval for the true p_i

$$|p_i - x_i| \leq \beta \sqrt{\frac{x_i (1 - x_i)}{K}} \quad (24)$$

for $i \in [1, n]$ (Theorem 11.6 of [12]).

□

Proof of Theorem 2

Proof:

From Theorem 1, we know that each random variable x_i for $i \in [1, n]$ has an approximate normal distribution $N(p_i, \frac{p_i(1-p_i)}{K})$ when the population K is such that $K p_i \geq 5$. Since $\hat{q}_i = c v_i x_i$, then \hat{q}_i also has an approximate normal distribution $N(c v_i p_i, \frac{c^2 v_i^2 p_i(1-p_i)}{K})$. Since \hat{q}_i would be very close to q_i when K is large, we can make the following approximation,

$$\hat{q}_i \ln \hat{q}_i \approx q_i \ln q_i + (1 + \ln q_i)(\hat{q}_i - q_i), \quad (25)$$

which is a linear function of the random variable \hat{q}_i . Therefore, $\hat{q}_i \ln \hat{q}_i$ is also approximate normal distribution,

$$E[\hat{q}_i \ln \hat{q}_i] = q_i \ln q_i, \quad (26)$$

$$\text{var}(\hat{q}_i \ln \hat{q}_i) = (1 + \ln q_i)^2 \frac{c^2 v_i^2 p_i(1-p_i)}{K}. \quad (27)$$

1. Since \hat{C}_e is a linear combination of $\hat{q}_i \ln \hat{q}_i$ for $i \in [1, n]$, \hat{C}_e also has an approximate normal distribution. Then, we have,

$$E[\hat{C}_e] = - \sum_{i=1}^n E[\hat{q}_i \ln \hat{q}_i] = - \sum_{i=1}^n q_i \ln q_i = C_e,$$

and

$$\text{var}(\hat{C}_e) = \sum_{i=1}^n \text{var}(\hat{q}_i \ln \hat{q}_i) + \sum_{i \neq j} \text{Cov}(\hat{q}_i \ln \hat{q}_i, \hat{q}_j \ln \hat{q}_j).$$

For calculating $\text{Cov}(\hat{q}_i \ln \hat{q}_i, \hat{q}_j \ln \hat{q}_j)$, recall equations (25) and (27),

$$\begin{aligned} \text{Cov}(\hat{q}_i \ln \hat{q}_i, \hat{q}_j \ln \hat{q}_j) &= E[(\hat{q}_i \ln \hat{q}_i - q_i \ln q_i)(\hat{q}_j \ln \hat{q}_j - q_j \ln q_j)] \\ &= (1 + \ln q_i)(1 + \ln q_j) \text{Cov}(\hat{q}_i, \hat{q}_j) \\ &= (1 + \ln q_i)(1 + \ln q_j) c^2 v_i v_j \text{Cov}(x_i, x_j). \end{aligned}$$

We know that for a multinomial distribution,

$$\text{Cov}(x_i, x_j) = -\frac{p_i p_j}{K}. \quad (28)$$

Therefore,

$$\text{Cov}(\hat{q}_i \ln \hat{q}_i, \hat{q}_j \ln \hat{q}_j) = -(1 + \ln q_i)(1 + \ln q_j) \frac{q_i q_j}{K}. \quad (29)$$

So, the variance of \hat{C}_e is

$$\begin{aligned} \text{var}(\hat{C}_e) &= - \sum_{i=1}^n \sum_{j=1}^n (1 + \ln q_i)(1 + \ln q_j) \frac{q_i q_j}{K} \\ &\quad + \sum_{i=1}^n (1 + \ln q_i)^2 \left(\frac{c v_i q_i}{K} - \frac{q_i^2}{K} \right) + \sum_{i=1}^n (1 + \ln q_i)^2 \frac{q_i^2}{K} \\ &= - \sum_{i=1}^n \sum_{j=1}^n (1 + \ln q_i)(1 + \ln q_j) \frac{q_i q_j}{K} + \sum_{i=1}^n (1 + \ln q_i)^2 \frac{c v_i q_i}{K} \\ &= - \frac{(1 - C_e)^2}{K} + \sum_{i=1}^n (1 + \ln q_i)^2 \frac{c v_i q_i}{K}. \end{aligned}$$

2. From equation (16), we see that this estimation is unbiased. Since $0 < v_i \leq 1$ and q_i is non-zero for each $i \in [1, n]$, and $var(\hat{C}_e) \propto \frac{1}{K}$ in equation (17), we can have

$$\lim_{K \rightarrow \infty} var(\hat{C}_e) = 0.$$

Therefore, we have

$$\lim_{K \rightarrow \infty} \hat{C}_e = C_e,$$

and hence \hat{C}_e is a consistent estimator of C_e .

3. If the population size is large enough that $K p_m \geq 5$, then with confidence of $(1 - \alpha)100\%$, the estimation error of the complexity \hat{C}_e is at most $\beta \sigma_{\hat{C}_e}$, where $\sigma_{\hat{C}_e}^2 = var(\hat{C}_e)$. That is,

$$|\hat{C}_e - C_e| \leq \beta \sigma_{\hat{C}_e}.$$

Replacing p_i with the approximation x_i in $\sigma_{\hat{C}_e}$ and considering $0 < v_i \leq 1$, we have

$$\begin{aligned} \sigma_{\hat{C}_e}^2 &\approx -\frac{(1 - \hat{C}_e)^2}{K} + \sum_{i=1}^n (1 + \ln \hat{q}_i)^2 \frac{c v_i \hat{q}_i}{K} \\ &\leq -\frac{(1 - \hat{C}_e)^2}{K} + \frac{c}{K} \sum_{i=1}^n (1 + \ln \hat{q}_i)^2 \hat{q}_i \\ &\leq -\frac{(1 - \hat{C}_e)^2}{K} + \frac{c}{K} (1 - 2\hat{C}_e - \ln \hat{q}_{min} \hat{C}_e) \\ &= -\frac{(1 - \hat{C}_e)^2}{K} + \frac{c}{K} (1 - (2 + \ln \hat{q}_{min}) \hat{C}_e) = \delta^2. \end{aligned}$$

Since $\sigma_{\hat{C}_e}^2 > 0$, it is always true that $\delta^2 \geq 0$. We have the third result of this theorem,

$$|\hat{C}_e - C_e| \leq \beta \delta.$$

□